# Meaning of the $\lambda$ parameter of skew–normal and log–skew normal distributions in fluid geochemistry

**Antonella Buccianti**

Department of Earth Sciences, Università degli Studi di Firenze
Firenze, Italy

**Mailing address:**
A. Buccianti
Dip. Scienze della Terra, Università degli Studi di Firenze
Via G. La Pira 4, 50125, Firenze, Italy
e-mail:antonella.buccianti@unifi.it
fax: +39-055-284571

**Keywords:** water chemistry, Vulcano island, probability models, logratios, skew–normal distributions family

# Abstract

The literature related to skew–normal distributions has grown rapidly in recent years but at the moment few applications concern the description of natural phenomena with this type of probability models, as well as the interpretation of their parameters. The skew–normal distributions family represents an extension of the normal family to which a parameter ($\lambda$) has been added to regulate the skewness. The development of this theoretical field has followed the general tendency in Statistics towards more flexible methods to represent features of the data, as adequately as possible, and to reduce unrealistic assumptions as the normality that underlies most methods of univariate and multivariate analysis. In this paper an investigation on the shape of the frequency distribution of the logratio $\ln(Cl^-/Na^+)$ whose components are related to waters composition for 26 wells, has been performed. Samples have been collected around the active center of Vulcano island (Aeolian archipelago, southern Italy) from 1977 up to now at time intervals of about six months. Data of the logratio have been tentatively modeled by evaluating the performance of the skew–normal model for each well. Values of the $\lambda$ parameter have been compared by considering temperature and spatial position of the sampling points. Preliminary results indicate that changes in $\lambda$ values can be related to the nature of environmental processes affecting the data.

# Introduction

According to Azzalini (1985) the skew–normal distribution is a family of distributions on the real line including the normal model, but with an extra parameter which allows the density to have positive or negative skewness. In the density function equation, depending from the expected value $\mu$ and variance $\sigma^2$, $\lambda \in \Re$ is the parameter that regulates the

shape of the distribution. Following this approach Mateu-Figueras (2003) has investigated the univariate logskew–normal distribution as a possible model for compositional data. In this framework, a positive random variable $X$, with support $\Re^+$, follows the univariate logskew–normal distribution, with parameters $\mu$, $\sigma^2$ and $\lambda$, if the transformed variable $Y = \ln(X)$ follows the skew–normal distribution. However, attention has to be posed in applying models defined on the whole real line to data which are actually constrained to an interval, like concentrations or similar quantities. These models might be a good approximation but the features of compositional data are not yet completely captured. With regard to this aspect, Mateu-Figueras et al. (2005) propose a further development with the application of the skew family models on logratios.

The chemical constituents of the waters are usually reported as concentrations in gravimetric units as milligrams per liter ($mg/l$ or $ppm$ if water density is equal to 1 $g/cm^3$) or milli equivalents per liter ($meq/l$). In both cases, for one component of the composition only a small part of the positive real line is used, thus compromising the application of classical statistical methodologies. Here, the logratio approach proposed by Aitchison allows us to model data with probability density functions in a correct way. The topic has been widely neglected, although the problem discussed in several papers (*i.e.* Ahrens, 1953;1954a,1954b;1957; Aubrey, 1954; 1956; Chayes, 1954; Miller and Goldberg, 1955; Vistelius, 1960). The goodness of the logratio transformation principle is based on the fact that there is a one–to–one correspondence between compositional vectors and associated log–ratios so that any statement about compositions can be reformulated in terms of log–ratios and vice versa. The transformation removes the problem of a constrained sample space (the unit simplex) opening up to all available standard statistical techniques. Moreover, the logarithmic function is monotonic increasing and consequently if a log–ratio increases, the ratio increases, an important aspect from the viewpoint of interpretation. The potentiality of the skew–normal distributions family, as *natural laws* to model logratios whose values are determined by the action of geochemical phenomena, is currently

under evaluation. In this paper an investigation on the shape of the frequency distribution of the logratio $\ln(Cl^-/Na^+)$, whose components are related to waters composition for 26 wells, has been performed. Samples have been collected around the active center of Vulcano island (Aeolian Archipelago, southern Italy) from 1977 up to now at time intervals of about six months. The logratios have been tentatively modelled by evaluating the performance of the skew–normal model for each well. Values of the $\lambda$ parameter obtained for each frequency distribution have been then compared by considering temperature and spatial position of the sampling points. Preliminary results indicate that some interesting features related to the processes affecting the data are associated with $\lambda$ changes and that this path of investigation may be full of promises.

## Geological and geochemical background

*Vulcano* is an island of a typical volcanic arc generated by subduction processes beneath the Tyrrhenian Sea, belonging to the Aeolian Archipelago (Sicily, southern Italy) and has had the latest eruption from 1888 to 1890. Since then, a fumarolic activity of varying intensity has continued up to now. Several years of geochemical investigations have proposed some models to describe the evolution of the fluids (water and gases) related to the volcanic system in time (*i.e.* Martini, 1980; 1989; 1996; Montalto, 1996; Capasso et al., 1999; 2001; Di Liberto et al., 2002). These studies allowed the identification of at least two aquifers, a shallow one of meteoric origin and a deeper one influenced by thermal activity. From a hydrological point of view, these aquifers are probably not physically separated. A current hypothesis considers that the shallower less saline aquifer floats over the more saline one of marine origin and affected by the interaction with volcanic fluids of deeper origin. In this framework, the differences observed in the wells are attributable to lateral permeability variations, local alteration processes, and/or to the presence of

areas of preferential up flow of volcanic fluids. In this type of environmental context the aggressive character of the water, which is able to mobilise the elements, is due to the input of carbon dioxide from the deep uprising gaseous flow into the aquifer. The presence of hydro magmatic deposits, which appear to have undergone early syn-depositional alteration processes, contributes elements from secondary minerals as calcium sulphate, calcium fluoride, sodium chloride and so on. According to their solubilities in aqueous solutions, chemical components can be leached away even if in presence of a weak alteration of surface waters. Systematic studies have been in progress since 1977 by the Geochemistry Units of the University of Florence. At present, 977 samples of ground waters, pertaining to 26 wells located in the area surrounding the active crater, have been sampled and analysed at regular time intervals by considering the concentrations ($ppm$) of $Ca^{2+}$, $Na^{+}$, $Mg^{2+}$, $K^{+}$, $HCO_3^{-}$, $SO_4^{2-}$ and $Cl^{-}$. From the point of view of fundamental chemical composition, the samples can be divided into two main groups, alkaline chloride and earth alkaline sulphate, and a small carbonatic subgroup. The differences of chemical characters correspond to diverse sampling sites where dissimilar geochemical processes appear to be dominant. In the area of *Porto di Levante* and under the main crater relatively deep phenomena of rock leaching, with possible contributions of hydrothermal solutions, are active. In the area of the *Porto di Ponente* the presence of a shallow brackish aquifer affects the chemical composition of waters. Finally, in an area located between the previous ones a mixing of different phenomena is the most important feature of the sampled waters (Martini, 1980). In this paper single logratios, whose components are properly chosen from a geochemical point of view (*i.e.* anions and/or cations with the same charge, species potentially derived from the same source and so on), have been modelled by probability distribution functions. The potentiality of the skew–normal distributions family, as a tool to describe natural phenomena in the univariate case, has been consequently probed. The parameters of the distributions have been estimated by using the maximum likelihood method, working with the routines of the Matlab version

software library available at http://azzalini.stat.unipd.it/SN/index.html.

## Methodological approach and results

The data pertaining to all the sampled wells have been investigated by considering the shape of the frequency distribution of $\ln(Cl^-/Na^+)$. Independently from the used measure unit, data to be managed are compositional and their components represent proportions of some whole (Buccianti and Pawlowsky-Glahn, 2005). If concentrations in *ppm* are converted in *meq/l*, as in our case, the ratio between $Cl^-$ and $Na^+$ is equal to 1 in rain water (halite dissolution) and tends to increase to 1.2 in sea–water. Consequently, shift in the values of the ratio with respect to 1 (0 in the case of logratio) can give interesting information about the geochemical processes contributing or subtracting ions. In general, if a river water sample contains 1.2 *mg/l* of $Na^+$ (or *ppm*) this corresponds to 1.2/22.99 (weight of the element)=0.052 $mmol\ Na^+/l$ and to 0.052 $mmol\ Na^+/l \times 1 = 0.052$ *meq/l* (1 is the charge of the cation).

After several years of research and discussion now we know that a suitable sample space for compositional data is the unit simplex and that data modelling can be performed under the logistic normal theory (Aitchison, 1982;1986). The theory is based on the transformation of compositional data from the simplex to the real space and modelling the transformed data by using the univariate or multivariate normal distribution. Consequently, the additive logratio transformation leads to the additive logistic normal model whose appealing properties have been widely discussed. However, following this path may be inappropriate when transformed data continue to present some skewness. By considering this, Azzalini (1985) and Azzalini and Dalla Valle (1996) have introduced as an alternative the skew normal distributions family in real space, a family that includes the normal one as a special case. Its most important feature is the presence of an extra

parameter able to manage the presence of skewness. The first step of investigation has been performed by considering the performance of the logistic skew–normal model for the $\ln(Cl^-/Na^+)$ frequency distribution in the sampled 26 wells. In order to apply the probability density functions in a correct way, presence of independent observations has to be hypothesised. Our data, instead, could be characterised by a time–dependence structure. In fact $x_{ij}$ is the response of the case $i$, $i = 1, \cdots, n$, at time j, $j = 1, \cdots, q$ and the $n$ cases are divided in $g$ groups (the wells). Consequently, for each group we have repeated measurements since the same spatial location has been monitored at several fixed occasions. Observations that are made at different times on the same experimental unit will always show some correlation. This correlation will not generally be predictable although it is reasonable to suppose that observations made close together in time, will be more highly correlated than ones taken far apart in time, as in the case of *Vulcano* data (Krzanowski and Marriott, 1995). Runs test (with respect to the median) confirms this hypothesis for the majority of the wells with $\alpha = 0.01$.

According to Azzalini (1985) a random variable $X$, with support the real line, is said to have a univariate skew–normal distribution with location and scale parameters $\mu$ and $\sigma$ if it is continuous with density function

$$P(x_i) = 2\phi(x_i; \mu, \sigma)\Phi\left(\frac{\lambda(x_i - \mu)}{\sigma}\right) \qquad x \in \Re, \tag{1}$$

where $\phi$ is the normal density function with expected value $\mu$ and variance $\sigma^2$, $\Phi$ the standard normal distribution function. The notation $X \sim SN(\mu, \sigma, \lambda)$ is used to indicate that the variable $X$ follows the skew-normal model. The parameter $\lambda \in \Re$ and controls the shape of the distribution; when $\lambda = 0$ the density $SN(\mu, \sigma, 0)$ is equal to the $N(\mu, \sigma)$, while when $\lambda \to \pm\infty$ a half-normal density is the result (Azzalini, 1985). Consequently, besides the mean and the variance, the skew-normal distribution depends from a skewness index $\gamma_1$ that varies in the interval (-0.995,+0.995) and when $\lambda \to \pm\infty$, $\gamma_1 \to \pm 0.995$. In

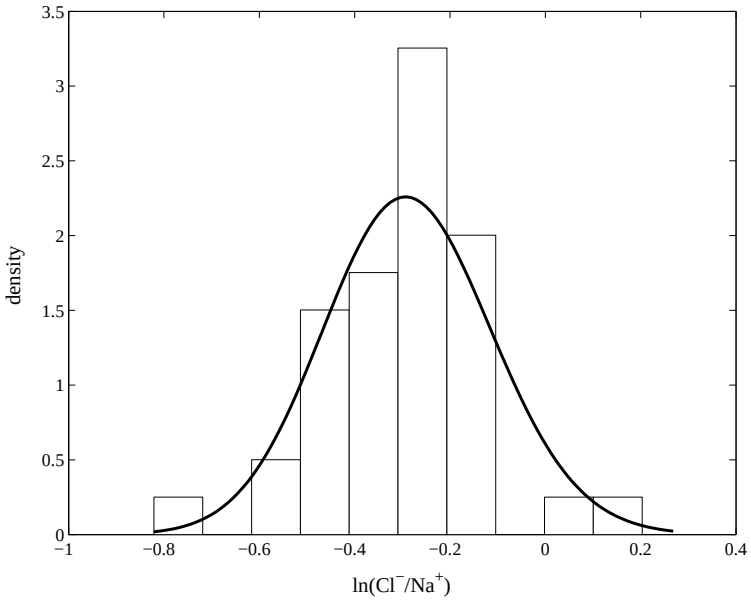


Figure 1: Histograms of $\ln(Cl^-/Na^+)$ and the fitted skew–normal model: well 1.

Table 1 the estimated values of the univariate skew–normal distribution for the logratio $\ln(Cl^-/Na^+)$ for each well (1,$\cdots$,26) are shown. The last column reports the significance of the normality test so that when this value is higher than 0.05 the hypothesis of normality of the logratio cannot be refused. There are 5 wells (3, 5, 6, 7 and 12) for which normal distribution is not the model to be used to describe the data for $\alpha = 0.05$. In all the other cases, even if the performance of the normal model is statistically acceptable, the presence of a light skewness makes the skew–normal model preferable. Examples of these situations are reported in the histograms of Figures 1 and 2. The goodness of the skew–normal model, compared with the normal one, can be appreciated visualising the data on probability plots as in Figures 3 and 4. Besides the graphical evaluation, statistical tests can be used to decide about the performance of the skew–normal model. Mateu–Figueras (2003) has recently developed various tests for a parent univariate skew–normal distribution based on the empirical distribution function. The idea underlying these tests is to measure the difference between the hypothesized distribution function (the

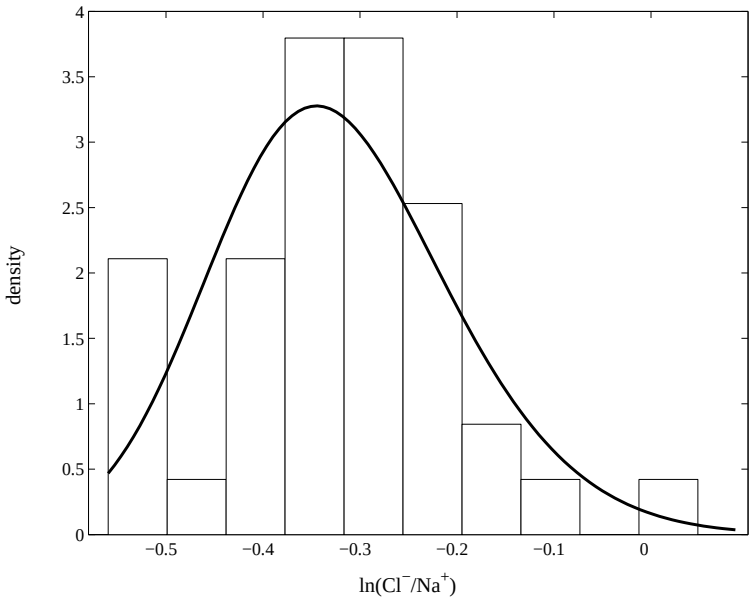


Figure 2: Histograms of $\ln(Cl^-/Na^+)$ and the fitted skew–normal model: well 2.

skew–normal) and the empirical distribution function computed from the sample. The difference can be measured by using various statistics as the Anderson–Darling ($A^2$), the Cramer–von Mises ($W^2$), the Watson ($U^2$), the Kolmogorov–Smirnov ($D$) and the Kuiper ($V$) ones, following the $\chi^2$ distribution under the null hypothesis. A first evaluation of the normal and skew–normal models has been obtained by considering the log–likelihood function values. Results for the $\ln(Cl^-/Na^+)$ reveal that skew–normal model is only slightly better than the normal one, with the exception of 5 wells for which a plurimodal distribution is present. Application of fit goodness test leads to the same conclusions for $\alpha = 0.01$. In Figure 5 an example of probability plots where both the normal and the skew–normal model are not able to represent the data is reported. The investigation of the frequency distribution of $\ln(Cl^-/Na^+)$ for most of the wells indicates that the normal and skew–normal models are both statistically valid to represent data. This result is not strange, because the normal model is a particular case of the skew–normal one and in several cases a low value of the skewness index $\gamma_1$ has been obtained. Values of $\lambda$ for wells

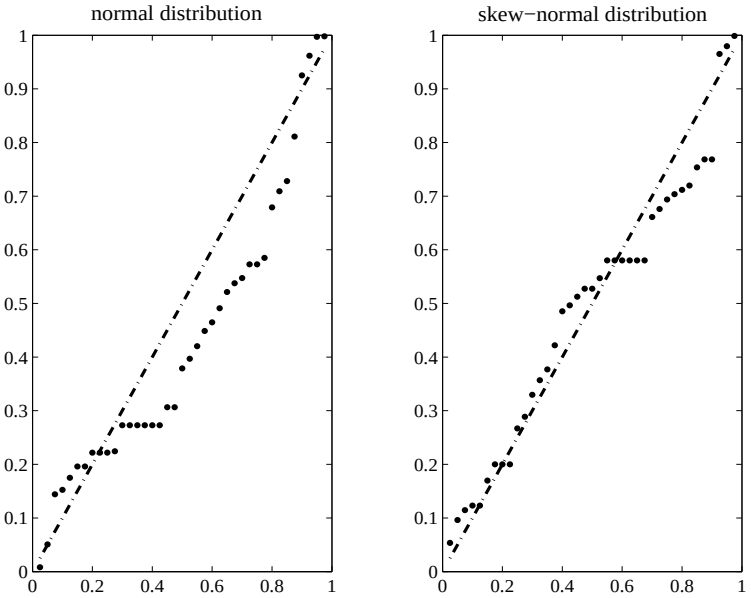


Figure 3: Probability plots of $\ln(Cl^-/Na^+)$: well 1.

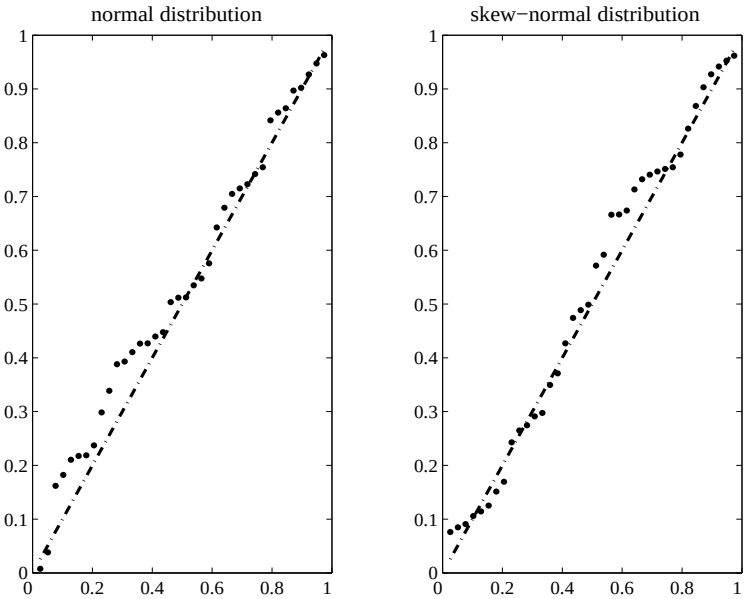


Figure 4: Probability plots of $\ln(Cl^-/Na^+)$: well 14.

Table 1: Estimated values of the univariate skew–normal distribution for the logratio $\ln(Cl^-/Na^+)$ for each well (1,$\cdots$,26). In the columns the mean and the standard deviation of the shew–normal model, followed by the values of $\lambda$, $\gamma_1$ and the significance of the normality test, are reported.

| Estimated values for each well (1,$\cdots$,26) | | | | | |
|---|---|---|---|---|---|
| | mean | std. dev. | $\lambda$ | $\gamma_1$ | p–value |
| 1 | -0.29 | 0.18 | 0.80 | 0.08 | 0.7417 |
| 2 | 0.32 | 0.12 | 1.53 | 0.31 | 0.2466 |
| 3 | -0.22 | 0.25 | 786 | 0.99 | 0.0000 |
| 4 | -0.07 | 0.17 | -2.12 | -0.49 | 1.0000 |
| 5 | -0.09 | 0.26 | -4.89 | -0.85 | 0.0028 |
| 6 | -0.09 | 0.17 | 3.75 | 0.76 | 0.0192 |
| 7 | 0.14 | 0.16 | 3.84 | 0.77 | 0.0088 |
| 8 | -0.60 | 0.17 | 2.54 | 0.58 | 1.0000 |
| 9 | -0.47 | 0.24 | -0.31 | -0.006 | 0.9836 |
| 10 | -0.25 | 0.31 | 0.72 | 0.62 | 1.0000 |
| 11 | -0.46 | 0.14 | -5.39 | -0.86 | 0.2463 |
| 12 | -0.45 | 0.19 | -7.97 | -0.94 | 0.0001 |
| 13 | -0.58 | 0.26 | 1.22 | 0.21 | 0.6832 |
| 14 | -0.36 | 0.21 | 4.54 | 0.83 | 0.1185 |
| 15 | -0.37 | 0.15 | 3.19 | 0.69 | 0.1535 |
| 16 | -0.48 | 0.18 | -0.98 | -0.13 | 0.6589 |
| 17 | -1.05 | 0.17 | 1.94 | 0.44 | 0.3388 |
| 18 | -0.56 | 0.35 | -2.34 | -0.54 | 1.0000 |
| 19 | -1.20 | 0.20 | 3.50 | 0.73 | 0.1023 |
| 20 | -0.81 | 0.25 | 2.20 | 0.51 | 1.0000 |
| 21 | -0.23 | 0.11 | 2.84 | 0.64 | 0.0972 |
| 22 | -0.55 | 0.17 | 8.61 | 0.94 | 0.1119 |
| 23 | -0.22 | 0.18 | -1.16 | -0.19 | 0.6326 |
| 24 | -1.86 | 0.31 | -1.58 | -0.33 | 0.4257 |
| 25 | -0.32 | 0.24 | 0.76 | 0.07 | 0.8255 |
| 26 | -0.39 | 0.35 | 2.94 | 0.66 | 1.0000 |

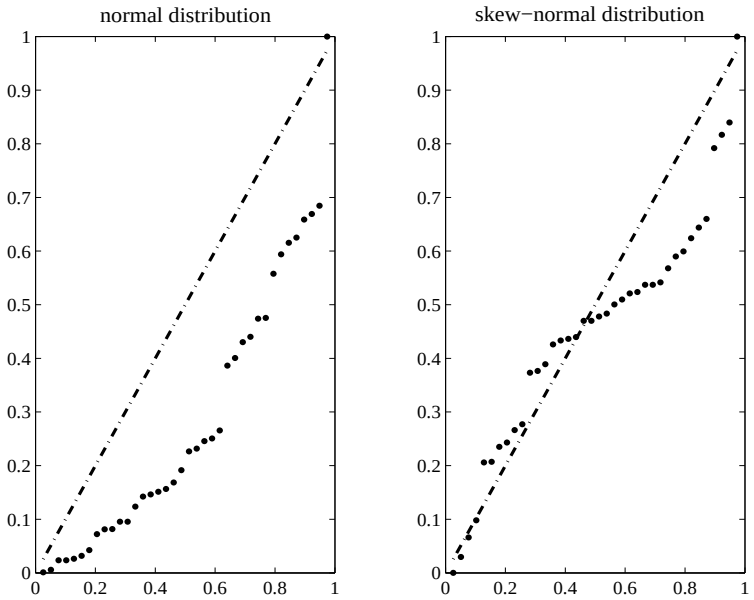


Figure 5: Probability plots of $\ln(Cl^-/Na^+)$: well 3.

of *Vulcano* island appear to cover a wide range corresponding to a 1) more or less good symmetry (normal model or skew–normal model with $\gamma_1 \to 0$), 2) presence of a moderate skewness, 3) presence of plurimodality (normal and skew–normal models not adequate). Consequently, a further investigation has been performed to verify possible relationships between $\lambda$ and external variables as temperature and spatial location. This in order to evaluate if processes able to shift distributions towards left or right are related to the shape parameter. If the $\lambda$ values are ordered by increasing temperature values, the obtained sequence is not random (run test with respect to median, $\alpha = 0.05$). The passage from negative to positive $\lambda$ values involves an increase in temperature values. Negative $\lambda$ values are related to $\ln(Cl^-/Na^+)$ more and more less negative implying a contribution of $Cl^-$ by sea–water, according with lower temperature too. On the other hand, positive $\lambda$ values are related to $\ln(Cl^-/Na^+)$ more and more negative implying a contribution of $Na^+$ by rock–weathering, a process facilitated by higher temperature. These considerations are also confirmed by the spatial position of the wells. In the first case most of them are

located in the area towards the Porto di Ponente, where the influence of a shallow brackish aquifer has been recognised in several investigations. In the second case most of them are located at the base of the active crater, an area in which relatively deep phenomena of rock leaching with contribution of hydrothermal solutions have been registered. In this context, depending on the depth of the well too, some complex situations are present in which mixing phenomena lead to plurimodal distributions.

## Concluding remarks

A discussion about the possible meaning of $\lambda$, a shape parameter controlling the skewness of the skew–family distribution models, has been presented. This research is aimed to try to describe natural phenomena with different models with respect to those usually adopted. As an application example waters composition collected at *Vulcano* island (Sicily, Italy) has been investigated. In particular, the shape of the frequency distribution of the log–ratio $\ln(Cl^-/Na^+)$, has been analysed by considering data from 26 wells. The performance of the normal and skew-normal models has been evaluated for each well; slight differences in the log–likelihood function are in favour of the skew–normal distribution. Values of $\lambda$ cover a wide range and their increase is related to an increase in temperature and to the sampling site. In this framework, extreme negative $\lambda$ values represent wells, mainly toward the Porto di Ponente area, in which contribution of sea–water ($Cl^-$) is a dominant process that can buffer the temperature values too. On the other hand, less negative $\lambda$ values represent wells, mainly at the base of the active crater, in which contribution of rock–weathering ($Na^+$) is a dominant process favoured by higher temperature values. Results obtained in this work indicate that $\lambda$ values of the skew–normal distributions family are potentially useful to characterise natural phenomena and further investigation has to be performed on other logratios.

## Acknowledgements

This research has been financially supported by Italian MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca Scientifica e Tecnologica), PRIN 2004, through the GEOBASI project (prot. 2004048813–002).

## References

Ahrens, L. H. (1953). A fundamental law of Geochemistry. *Nature 172*, 1148.

Ahrens, L. H. (1954a). The lognormal distribution of the elements (a fundamental law of Geochemistry and its subsidiary. *Geochimica and Cosmochimica Acta 6*, 49–74.

Ahrens, L. H. (1954b). The lognormal distribution of the elements. II. *Geochimica and Cosmochimica Acta 6*, 121–132.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society Series B 44(2)*, 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London. 463 p.

Aubrey, K. V. (1954). Frequency distribution of the concentrations of the elements in rocks. *Nature 174*, 141–142.

Aubrey, K. V. (1956). Frequency distributions of elements in igneous rocks. *Geochimica and Cosmochimica Acta 9*, 83–90.

Azzalini, A. (1985). A class of distribution which includes the normal ones. *Scandinavian Journal of Statistics 12*, 171–178.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika 83(4)*, 715–726.

Buccianti, A. and V. Pawlowsky-Glahn (2005). New perspectives on water chemistry and compositional data analysis. *Mathematical Gology 37(7)*, 707–731.

Capasso, G., W. D'Alessandro, R. Favara, S. Inguaggiato, and F. Parello (2001). Interaction between the deep fluids and the shallow groundwaters on Vulcano Island (Italy). *Journal of Volcanology and Geothermal Research 108*, 187–198.

Capasso, G., R. Favara, S. Fracofonte, and S. Inguaggiato (1999). Chemical and isotopic variations in fumarolic discharge and thermal waters at Vulcano Island (Aeolian Islands, Italy) during 1996: evidence of resumed volcanic activity. *Journal of Volcanology and Geothermal Research 88*, 167–175.

Chayes, F. (1954). The lognormal distribution of elements: a discussion. *Geochimica and Cosmochimica Acta 6*, 119–121.

Di Liberto, V., P. M. Nuccio, and A. Paonita (2002). Genesis of chlorine and sulphur in fumarolic emissions at Vulcano Island (Italy): assessment of pH and redox conditions in the hydrothermal system. *Journal of Volcanology and Geothermal Research 116*, 137–150.

H., A. L. (1957). Lognormal type distribution. III. *Geochimica and Cosmochimica Acta 11*, 205–213.

Krzanowski, W. and F. H. C. Marriott (1995). *Multivariate Analysis, part II.* Kendall's Library of Statistics 2, Arnold, London. 376 p.

Martini, M. (1980). Geochemical survey on the phreatic waters of Vulcano (Aeolian Islands, Italy). *Bulletin of Volcanology 43(1)*, 265–274.

Martini, M. (1989). The forecasting significance of chemical indicators in areas of quiescent volcanism: examples from Vulcano and Phlegrean Fields (Italy). *Volcanic Hazard, IAVCEI Proceedings in Volcanology 1, J. H. Latter(Ed.), Springer–Verlag, Berlin Heidelberg, Germany*, 372–383.

Martini, M. (1996). Chemical characters of the gaseous phase in different stages of volcanism: precursors and volcanic activity. *Monitoring and Mitigation of Volcanic Hazard, Scarpa/Tilling (Eds.), Springer–Verlag, Berlin Heidelberg, Germany*, 200–219.

Mateu-Figueras, G. (2003). *Models de distribució sobre el símplex*. Ph. D. Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2005). The additive logistic skew-normal distribution on the simplex. *Mathematical Geology, submitted*.

Miller, R. L. and E. D. Goldberg (1955). The normal distribution in Geochemistry. *Geochimica and Cosmochimica Acta 8*, 53–62.

Montalto, A. (1996). Signs of potential renewal of eruptive activity at La Fossa (Vulcano, Aeolian Islands). *Bulletin of Volcanology 57*, 483–492.

Vistelius, A. B. (1960). The shew frequency distributions and the fundamental law of the geochemical processes. *Journal of Geology 68*, 1–22.